# ORIN-Lyrics: A Multilingual Nigerian Song Lyrics Dataset and Baseline for Efficient Language Detection

**Sakinat O. Folorunso** sakinat.folorunso@oouagoiwoye.edu.ng
**Oluwagbenga Odunsi** odunsioluwagbenga5@gmail.com
**Ayodele David** ayodeleifeoluwa600@gmail.com
**Daniel Olaleye** danielolaleye080@gmail.com
**Fatimah Salami** fatimah.salami@gmail.com
**Oluwakemi Giwa** giwa.oluwakemi@oougoiwoye.edu.ng
*The Artificial Intelligence Research Group,*
*Department of Computer Science,*
*Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria*

**Editors:** Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

## Abstract

Natural language processing (NLP) needs more research on African languages because current studies fail to capture cultural musical domains. This study presents ORIN-Lyrics, a multilingual dataset containing Nigerian song lyrics that follows FAIR data principles (Findable, Accessible, Interoperable, and Reusable) for multilingual NLP research and cultural AI research. The dataset includes 853 songs representing 22 musical genres and 18 language categories, featuring Yoruba, English, Nigerian Pidgin, and mixed language compositions with code-switching. The dataset contains both lyrics and structured metadata, including artist names, genre labels, and language annotations. The corpus contains 124,801 tokens and 12,098 unique words, demonstrating the linguistic diversity found in Nigerian music. An embedding based visualisation method displays distinct semantic groups between different language groups. The dataset demonstrates its practical value through a baseline genre classification experiment using Term Frequency-Inverse Document Frequency (TF-IDF) features and multi-class logistic regression, achieving an accuracy of 0.54 — substantially exceeding the random baseline of approximately 4.5enabling the creation of African language technologies that match local needs

**Keywords:** Multilingual NLP, African languages, Nigerian music lyrics, Code-switching, Cultural AI, Low-resource languages, Dataset curation

## 1. Introduction

### 1.1. Context: The NLP Resource Divide

Natural Language Processing (NLP) has advanced rapidly in recent years because deep learning technology, large datasets, and multilingual language models have become widely available. These developments have enhanced machine translation, information retrieval, and conversational systems. Modern NLP technologies provide benefits to various languages throughout the world; yet their advantages remain inaccessible to many. Most datasets and pretrained models are developed for a small group of high-resource languages such as English, Chinese, and Spanish, which results in African languages being excluded from computational linguistics research. The African continent contains more than 2,000 languages,

yet this extreme linguistic diversity creates a research imbalance. The majority of these languages lack sufficient digital resources to enable extensive NLP research. The development of language technologies for African communities faces obstacles because of a shortage of publicly accessible corpora, linguistic annotations, and benchmark datasets (Orife et al., 2020; Nekoto et al., 2020; Folorunso et al., 2022) Recent initiatives have begun to expand NLP resources for African languages. Collaborative projects such as Masakhane (Nekoto et al., 2020), MasakhaNER (Adelani et al., 2021), AfriBERTa (Ogueji, 2022). AfroLM (Dossou et al., 2022), and the No Language Left Behind (NLLB) project (Costa-Jussà et al., 2022) have made significant progress toward improving language technologies for underrepresented languages. Existing datasets, however, demonstrate a primary focus on formal textual domains, including translation corpora, news articles, and web documents.

### 1.2. The Gap: Moving Beyond Formal Text Domains

The existing initiatives have made major progress in African NLP research, but there exists a shortage of language datasets that reflect African cultural heritage. Music lyrics represent a particularly rich domain where language, culture, and identity intersect. Nigerian music — spanning genres such as Afrobeat, Afropop, Gospel, Fuji, and contemporary Hip-Hop — often blends multiple languages within a single song. Nigerian lyrics demonstrate a unique character through artists who code-switch between Yoruba, English, and Nigerian Pidgin. These language patterns mirror everyday communicative norms in Nigerian society. Despite its global impact, Nigerian music remains an underdeveloped area for NLP research (Folorunso et al., 2025). The existing lyric datasets mainly focus on Western music, which assumes that artists sing only in one language. (Hung et al., 2022) established music lyrics as linguistic resources for research purposes. However, researchers have not yet created a comprehensive multilingual database enabling investigation of Nigerian music lyrics at a large scale. This study directly addresses that gap through the ORIN-Lyrics dataset, providing the first large-scale, FAIR-compliant corpus of Nigerian song lyrics with structured multilingual metadata.

### 1.3. Proposed Solution and Contributions

To address this problem, this study presents ORIN-Lyrics: a multilingual dataset containing Nigerian song lyrics designed to support research in multilingual natural language processing and cultural artificial intelligence. The dataset contains 853 songs covering 22 musical genres and 18 language categories, including Yoruba, English, Nigerian Pidgin, and mixed-language compositions. The dataset contains lyrical text together with structured metadata — including artist names, genre labels, and language annotations — enabling various linguistic and computational studies. As shown in Table 1, existing African NLP resources concentrate mainly on translation and formal text corpora, whereas ORIN-Lyrics documents the multilinguallanguage usage found in Nigerian music

### 1.4. Contributions of This Work

This paper establishes three main contributions:

**Dataset Curation.** The ORIN-Lyrics dataset presents a multilingual collection of Nigerian song lyrics comprising 853 songs from 22 musical genres and 18 language categories, curated in accordance with FAIR data principles (Wilkinson et al., 2016; Folorunso et al., 2022) .

Table 1: Comparison of ORIN-Lyrics with existing African NLP datasets.

| Dataset | Domain | Languages | Dataset Size | Code-Switching | Cultural Metadata |
|---|---|---|---|---|---|
| Masakhane MT Corpus (Nekoto et al., 2020) | Machine Translation | 30+ African languages | Large parallel corpus | No | No |
| MasakhaNER (Adelani et al., 2021) | Named Entity Recognition | 10 African languages | 20K+ sentences | No | No |
| AfriBERTa (Ogueji, 2022) | Language Model Corpus | 11 African languages | Large web corpus | Limited | No |
| AfroLM (Dossou et al., 2022) | Multilingual Language Model | Multiple African languages | Large multilingual corpus | Limited | No |
| NLLB Dataset (Costa-Jussà et al., 2022) | Machine Translation | 200+ languages | Large multilingual corpus | No | No |
| ORIN-Lyrics (This study) | Nigerian Music Lyrics | Yoruba, English, Pidgin | 853 songs lyrics | Yes | Yes |

**Empirical Analysis**. The dataset analysis includes vocabulary statistics, multilingual language distributions, code-switching patterns, and embedding-based visualisations that display the linguistic structure of Nigerian music lyrics.

**Predictive Baselining**. The dataset establishes baseline genre classification experiments, demonstrating that it contains valuable predictive information for downstream NLP tasks.

Table 1 shows that the majority of African NLP datasets focus on machine translation, named entity recognition, and multilingual language modelling from formal text sources. The ORIN-Lyrics dataset differs from these existing resources by capturing informal multilingual creative expression, code-switching, and struc tured cultural metadata

## 2. Methodology

This section describes the process used to construct the ORIN-Lyrics dataset and to perform baseline analyses of the corpus. The methodology covers data collection, metadata extraction, text preprocessing, feature representation, multilingual analysis, and baseline classification experiments.

### 2.1. Dataset Construction Pipeline

The ORIN-Lyrics dataset (Folorunso et al., 2025) was created through a structured data collection and curation pipeline designed to capture multilingual Nigerian song lyrics while preserving their linguistic and cultural context. The pipeline consists of five stages: lyrics collection, metadata extraction, data cleaning and normalisation, tokenisation, and final dataset compilation. Figure 1 illustrates the full workflow. Lyrics and associated metadata were collected from publicly available lyrics repositories and music platforms using automated web scraping scripts implemented in Python. Web scraping is widely used in

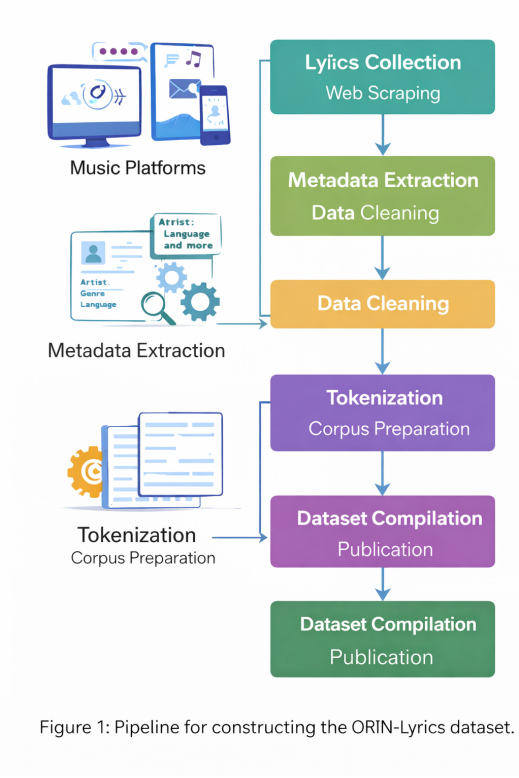



Figure 1: Pipeline for constructing the ORIN-Lyrics dataset.

corpus construction because it allows scalable extraction of textual data from online sources Mitchell (2024). After automated collection, the dataset underwent manual verification to ensure consistency. Duplicate records were eliminated and metadata consistency was enforced. Songs with incomplete metadata or corrupted text were excluded. The final dataset includes 853 Nigerian songs representing 22 musical genres and 18 language categories. The construction process follows best practices for multilingual corpus development in African NLP research (Nekoto et al., 2020).

### 2.2. Metadata Extraction and Dataset Structure

The ORIN-Lyrics dataset contains structured records for each song, comprising its complete lyrical content and accompanying descriptive metadata. The attributes are summarized in Table 2. These include artist name, song title, album details, genre classification, language category, year of release, cultural origin, and the original source URL of the lyrics. The inclusion of the release year provides important temporal context for tracking evolving language use and genre trends in Nigerian music. The dataset's structured metadata enables support for various downstream tasks, including music information retrieval, genre classification, and cultural analytics. Prior research has shown that combining textual data with structured metadata improves interpretability and enables richer linguistic analysis (Roxbergh, 2019) The metadata attributes included in the dataset are summarized in Table 2.

Table 2: Metadata fields in the ORIN-Lyrics dataset

| Field | Description |
|---|---|
| Artist | Performing artist |
| Title | Song title |
| Album | Album or project |
| Genre | Musical genre |
| Language | Primary language |
| Lyrics | Raw lyrical text |
| Region | Cultural origin |
| Source URL | Link to lyrics source |

### 2.3. Text Preprocessing and Tokenization

A standard NLP preprocessing pipeline was applied to the collected lyrics before analysis. The processinvolved four steps: (1) converting all letters to lowercase, (2) eliminating punctuation and special characters,(3) normalising multiple spaces into a single space, and (4) dividing the text into separate word units (Jurafsky and Martin, 2026). These steps produce uniform text forms suitable for machine learning systems. Whitespace tokenisation maintains the multilingual character of Nigerian music lyrics, through which artists switch between Yoruba, English, and Nigerian Pidgin. The system represents each lyric document as a token sequence:

$$D = \{w_1, w_2, w_3, ..., w_n\} \tag{1}$$

where $D$ denotes a lyric document and$w_i$ represents the $i^{th}$ token. The complete corpus is represented as:

$$C = \{D_1, D_2, ..., D_m\} \tag{2}$$

(2) containing m total songs from the dataset.

### 2.4. Feature Representation

Term Frequency-Inverse Document Frequency (TF-IDF) was used to transform the song lyrics into numerical feature vectors for computational analysis. TF-IDF measures the importance of a word in a document relative to its occurrence across the corpus and is widely used in information retrieval and text classification (Salton and Buckley, 1988; Roxbergh, 2019). The TF-IDF score for term $t$ in document $d$ is defined as:

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t) \tag{3}$$

$$TF(t, d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \tag{4}$$

$$IDF(t) = \log \frac{N}{n_t} \tag{5}$$

where $f_{t,d}$ is the frequency of term $t$ in document $d$, $N$ is the total number of documents in the corpus, and $n_t$ is the number of documents containing term $t$.

### 2.5. Code-Switching Detection

Nigerian music uses multiple languages as a central lyrical feature. Artists frequently switch between Yoruba, English, and Nigerian Pidgin within a single song. Code-switching describes this linguistic mixing, which enables speakers to alternate between languages during their discourse Myers-Scotton (1993). Language labels assigned to each song were used to determine how often languages changed across the corpus. The transition count between songs with language labels $Li$ and $Lj$ is defined as:

$$T(L_i, L_j) = T(L_i, L_j) + 1 \tag{6}$$

A heatmap was used to visualise transition frequencies and detect multilingual patterns across the dataset. It is acknowledged that song-level language labels provide an approximation of code-switching patterns. Future work will introduce token-level or line-level annotations to enable finer-grained analysis of within song language alternation

### 2.6. Multilingual Sentence Embeddings

A pretrained model from the SentenceTransformers framework (all-MiniLM-L6-v2) was used to create sentence embeddings for studying semantic relationships between lyrics written in various languages. Sentence Transformers extend transformer-based language models such as BERT to produce dense semantic representations suitable for similarity analysis and clustering (Reimers and Gurevych, 2019). The embedding model transforms lyric document $D$ into a vector representation

$$e_D = f(D) \tag{7}$$

where $f(\cdot)$ is the embedding function

### 2.7. Dimensionality Reduction Using UMAP

The Uniform Manifold Approximation and Projection (UMAP) method was used to visualise the embedding space structure. UMAP provides a nonlinear dimensionality reduction method that preserves local neighbourhood structure when transforming high-dimensional data into a low-dimensional space suitable for visual representation (Healy and McInnes, 2024). The embedding set $\mathrm{E} = \{\mathrm{e}_1, e_2, ..., e_n\} is projected into a two-dimensional space via f : \mathbb{R}^k \rightarrow \mathbb{R}^2 preserving local relationships between documents$

### 2.8. Baseline Genre Classification

A baseline genre classification experiment was conducted to demonstrate how the dataset supports machinelearning tasks. Lyrics were transformed into TF-IDF vectors as defined in Equation (3). Multi-class logistic regression — a well-established probabilistic classifier for text categorisation (Hosmer Jr et al., 2013) — was employed using a one-vs-rest (OvR) strategy, enabling classification across all 22 genres. The probability that a document belongs to genre g is modelled as:

$$P(g|x_d) = \frac{1}{1 + e^{-(w^T x_d + b)}} \tag{8}$$

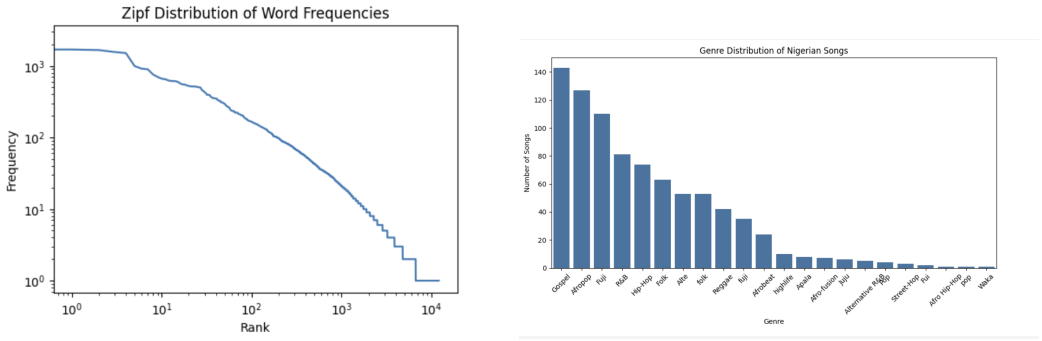



Figure 2: Lexical and genre characteristics of the ORIN-Lyrics dataset. Left: Zipf distribution of word frequencies. Right: Genre distribution.

where $w$ represents model weights and $b$ is the bias term. A stratified train-test split was used to partition the dataset, and classification accuracy was used as the primary assessment metric

## 3. Results and Analysis

The ORIN-Lyrics dataset is empirically evaluated through corpus statistics, lexical features, multilingual language distribution, and machine learning baselines.

### 3.1. Dataset Statistics

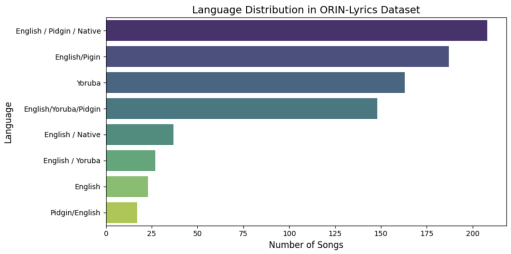
The ORIN-Lyrics dataset consists of 853 Nigerian songs covering 22 musical genres and 18 distinct linguistic categories. The corpus contains 124,801 tokens, including 12,098 distinct vocabulary terms. Each song contains approximately 146 tokens on average. The lyrics display the multilingual character of Nigerian music through the use of Yoruba, English, Nigerian Pidgin, and various mixed-language patterns.

### 3.2. Lexical Characteristics and Genre Distribution

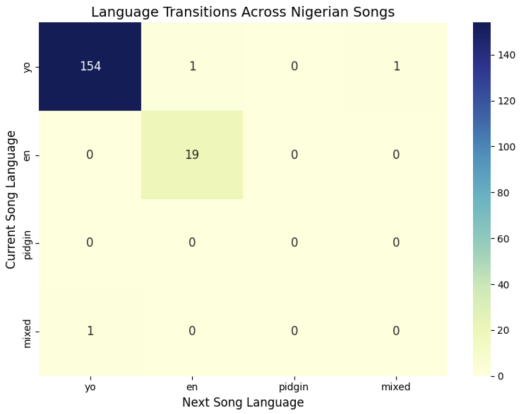
The lexical diversity of the corpus was assessed by examining vocabulary statistics and word usage patterns. The most common words — dey, love, know, like, and baby — reflect dominant themes of romantic relationships, emotional experiences, and social activities in Nigerian popular music. The corpus word frequency follows a Zipfian distribution, representing a typical pattern in natural language datasets. Figure 2 (left) shows the log–log frequency plot demonstrating how a small number of words dominate usage while most words appear rarely. Figure 2 (right) shows the distribution of songs across 22 musical genres. Gospel, Afropop, and Fuji demonstrate higher song counts, reflecting current trends in Nigerian music production and cultural expression.

### 3.3. Language Distribution and Code-Switching Patterns

Nigerian music often blends multiple languages within the same lyrics. The ORIN-Lyrics dataset reflects this multilingual style, containing songs in Yoruba, English, Nigerian Pidgin, and mixed-language forms. The distribution of language categories across songs shows that Yoruba, English, and Nigerian Pidgin are the most common, while English–Pidgin and English–Yoruba–Pidgin combinations represent the next most frequent categories. Figure 3 shows the language distribution across the corpus (left) and the song-level language tran-




(a) Language distribution




(b) Code-switching heatmap

Figure 3: Multilingual characteristics of the dataset.

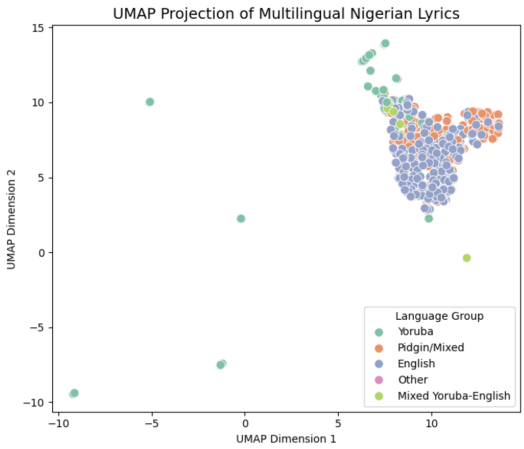



Figure 4: UMAP projection of multilingual sentence embeddings.

sition heatmap (right). The heatmap reflects the dataset's composition: Yoruba-labelled songs constitute the largest category, with a smaller number of transitions to mixed and English-labelled songs. This pattern is consistent with the song-level annotation granularity. Token-level annotation — identified as future work — will enable a more granular transition analysis.

### 3.4. Multilingual Embedding Visualization

The SentenceTransformer model (all-MiniLM-L6-v2) was used to create sentence embeddings, which were analysed through UMAP to discover semantic relationships between different language groups (Figure 4). The visualisation demonstrates distinct linguistic clusters: the primary grouping centres on Yoruba lyrics, while mixed-language compositions appear between the distinct language group clusters. Songs form groupings based on their linguistic similarities, while mixed-language lyrics position themselves between these groupings

### 3.5. Baseline Genre Classification

The dataset's value for machine learning applications was assessed through a baseline genre classification experiment using TF-IDF features and multi-class logistic regression (one-vs-rest). The model achieved an overall accuracy of 0.54 across 22 genres. The random baseline for a 22-class classification problem is approximately 4.5Error analysis shows that most misclassifications occur between stylistically related genres such as Afrobeat and Hip-Hop, which share linguistic elements and thematic patterns. Per-genre accuracy varies with dataset representation: genres with more songs in the training set achieve higher accuracy. Future work should report per-genre precision, recall, and F1-scores to provide a fuller picture of model performance across all 22 classes.

## 4. Discussion

The findings from Section 3 show that the ORIN-Lyrics dataset can support the investigation of multilingual language patterns in Nigerian music. This dataset captures linguistic features used in natural, creative, and culturally specific speech, whereas most African NLP datasets present only formal text and translation tasks. Recent efforts have built up NLP resources for African languages. The Masakhane project established a participatory research framework by creating machine translation datasets covering multiple African languages (Nekoto et al., 2020). The NLLB project created multilingual translation models supporting over 200 languages including multiple African languages (Costa-Jussà et al., 2022) Researchers have also developed various datasets and models for African languages, including the Yorùbá translation corpus (Alabi et al., 2020), MasakhaNER for multilingual named entity recognition (Adelani et al., 2021), and the African language models AfriBERTa (Ogueji, 2022) and AfroLM (Dossou et al., 2022). Most African NLP datasets continue to rely on formal text sources such as news articles and translation corpora. Music lyrics provide a culturally rich environment where language is used creatively to express identity, emotion, and social meaning. The AfroLyrics research established African music lyrics as valuable linguistic resources (Orife et al., 2020).TheORIN-Lyricsdataset extends this line of research by offering a curated multilingual corpus that captures code-switching patterns while providing structured cultural metadata, including dataset licensing information and clear citation instructions for downstream users.

## 5. Limitations and Ethical Considerations

### 5.1. Dataset Scope and Representativeness

The ORIN-Lyrics dataset does not fully represent the entire Nigerian music ecosystem. The corpus contains songs sourced from online public platforms, which means that traditional music existing only in physical form and not yet digitised is absent. The dataset also requires expansion to include more of Nigeria's approximately 500 indigenous languages and a broader range of musical genres — this constitutes a vital pathway for future work.

### 5.2. Class Imbalance

The dataset exhibits natural imbalances in both genre and language representation. Gospel, Afropop, and Fuji genres appear more frequently, while Yoruba and English dominate over less commonly spoken Nigerian languages. Although this reflects actual Nigerian music production patterns, it can introduce biases in machine learning model training.

Researchers should therefore consider appropriate sampling or balancing strategies when conducting supervised learning experiments.

### 5.3. Language Annotation Granularity

Language labelling presents a challenge for this dataset. Nigerian music demonstrates code-switching across multiple languages within a single song, which prevents a single language designation from fully describing all lyrics. The current dataset provides song-level language labels. Future work will address this limitation by introducing token-level or line-level language annotations that better represent multilingual patterns found in lyrical text.

### 5.4. Ethical and Legal Considerations

The ethical implications of working with cultural data require careful consideration. The dataset contains copyrighted lyrics, protected under copyright law; the dataset is intended for academic research use only. Where direct redistribution of copyright material is not permitted, the dataset references public sources rather than reproducing lyrics in full. Researchers must engage with this data with cultural sensitivity, avoiding false assumptions about slang terms and culturally specific social references in Nigerian music. With respect to web data collection, the lyrics were gathered from publicly accessible platforms. However, public accessibility does not automatically imply permission for scraping or redistribution. The construction of this dataset followed responsible data collection practices, including respect for the terms of service of source platforms and their robots.txt policies. The exact source URLs for each song are documented in the dataset metadata to support transparency and reproducibility. The dataset is released with an explicit open licence for academic use, along with clear citation instructions, to facilitate responsible downstream use. These considerations align with the dataset's compliance with FAIR data principles.

### 5.5. Threats to Validity

Several threats to the validity of this study are acknowledged. With respect to internal validity, the manual verification step helps ensure data quality, but residual errors in lyrics text or metadata may remain. Labelling consistency across 18 language categories was maintained through systematic checks, but the absence of interannotator agreement metrics is a limitation. With respect to external validity, the dataset reflects songs available on public platforms and may not generalise to the full diversity of Nigerian musical expression, particularly traditional and oral music traditions. The baseline genre classification result of 0.54 should be interpreted in context: it is substantially above the random baseline but leaves significant room for improvement with more sophisticated models and richer features.

## 6. Conclusion

This study developed ORIN-Lyrics as a multilingual dataset containing Nigerian song lyrics to support two fields of research: multilingual NLP and cultural artificial intelligence. The dataset contains 853 songs representing 22 musical genres and 18 language categories, comprising Yoruba, English, Nigerian Pidgin, and mixed-language songs. The dataset analysis demonstrates that Nigerian music linguistic patterns are evident through Zipfian lexical distributions, multilingual language usage, and code-switching behaviour. The UMAP embedding visualisations demonstrate how different language groups develop their own seman-

tic structure, while the baseline genre classification experiments confirm that lyrical features provide valuable information to support machine learning tasks. ORIN-Lyrics provides a resource combining lyrical text and structured cultural metadata, enabling researchers to study African music multilingual communication from a cultural perspective. The dataset will support investigations into multilingual NLP, cultural analytics, and African language technology development. Future work will expand the dataset through the addition of new languages and genre types, the creation of detailed token-level multilingual language annotations, and the reporting of comprehensive per-genre classification metrics. Increase of genre types and the creation of detailed multilingual language documentation

## 7. Data Availability

The study dataset is publicly available for academic research and reproducibility purposes. The dataset is currently hosted on Mendeley Data (Folorunso et al., 2025). To maintain double-blind peer review integrity, the direct repository URL has been redacted.

## References

[1] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[2] Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David I. Adelani, and Cristina España-Bonet. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.335/.

[3] Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[4] Bonaventure FP Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, 2022.

[5] Sakinat Folorunso, Ezekiel Ogundepo, Mariam Basajja, Joseph Awotunde, Abdullahi Kawu, Francisca Oladipo, and Abdullahi Ibrahim. Fair machine learning model pipeline implementation of covid-19 data. *Data Intelligence*, 4(4):971–990, 2022.

[6] Sakinat Folorunso, Tosin Akerele, Francisca Oladipo, and Oluwakemi Giwa. Orin lyrics dataset: A comprehensive corpus of multilingual nigerian song lyrics for nlp, 2025. URL https://doi.org/10.17632/2jfc8bjwwb.1.

[7] Sakinat O Folorunso, Sulaimon A Afolabi, and Adeoye B Owodeyi. Dissecting the genre of nigerian music with machine learning models. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6266–6279, 2022.

[8] John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, 2024.

[9] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[10] Yun-Ning Hung, Chih-Wei Wu, Iroro Orife, A. J. Hipple, William Wolcott, and Alexander Lerch. A large tv dataset for speech and music activity detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022, 2022. URL https://api.semanticscholar.org/CorpusID:252052082.

[11] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2026. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released January 6, 2026.

[12] Ryan Mitchell. *Web scraping with python*. " O'Reilly Media, Inc.", 2024.

[13] Carol Myers-Scotton. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press, 1993.

[14] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, 2020.

[15] Kelechi Ogueji. Afriberta: Towards viable multilingual language models for low-resource languages. 2022.

[16] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020.

[17] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[18] Linus Roxbergh. Language classification of music using metadata, 2019. ISSN 1650-8319.

[19] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[20] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan Willem Boiten, Luiz Bonino Da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship: Comment. *Scientific data*, 3:160018, 2016.